DOCUMENT RESUME

ED 236 251 TM 830 788

AUTHOR	Martin, Michael T.
TITLE	Arizona Teacher Proficiency Examination of the Basic Skills. A Special Report.
INSTITUTION	Arizona Tax Research Association, Phoenix.
PUB DATE	Oct 83
NOTE	25p.
AVAILABLE FROM	Arizona Tax Research Association, 1814 W. Washington St., Phoenix, AZ 85007 ($2.00).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Correlation; Elementary Secondary Education; Item Analysis; *Screening Tests; *State Programs; *Teacher Evaluation; *Testing Programs; Test Items; *Test Validity
IDENTIFIERS	*Arizona Teacher Proficiency Examination

ABSTRACT

The Arizona Tax Research Association analyzed the Arizona Teacher Proficiency Examination (ATPE). An item analysis of the 150 questions administered to 2,430 applicants from July 1982 through March 1983 was undertaken. A minimum point biserial correlation coefficient of .33 indicated the validity of an item. Nearly two-thirds (64 percent) of the items on the Reading section of the ATPE appear to be invalid, while 32 percent of the questions on the Math and 20 percent of the questions on the Grammar sections appear invalid. The discriminant validity of the ATPE is seriously questioned by this finding. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Morris

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE ATPE

The Arizona Tax Research Association (ATRA) analyzed the Arizona Teacher's Proficiency Examination (ATPE) using the ATRA computer and data purchased from the Arizona Department of Education. The analysis indicates the ATPE does not validly screen the applicants for Arizona teaching certificates, partly because the ATPE is not a valid test of basic skills, and partly because it is graded as a whole test rather than component parts. The analysis indicates that persons who pass the test may not be proficient in the basic skills, and that some persons who fail the test could very well be more proficient than some who pass.

Although the current test should not be used for screening, ATRA does not intend any criticism of using a valid test for screening teaching applicants. An adequate knowledge of basic skills is clearly a legitimate requirement for receiving a teaching certificate, but the decision to deny or award certification should be based on an accurate determination of whether the person has an adequate knowledge of basic skills. The current ATPE does not measure basic skills even by its own criteria, and the level of adequate knowledge was never determined. Not only do many of the questions on the ATPE appear invalid, but the test is scored for overall knowledge rather than knowledge on the individual basic skills, and the cutoff score of 80% was arbitrarily chosen for political reasons.

To analyze the basic skills test, ATRA paid the Arizona Department of Education to prepare a special computer printout of statistics for each

question on the ATPE without revealing the actual questions or answers. ATRA then evaluated each question by using the technically and mathematically sophisticated statistical measure that is used to validate Scholastic Aptitude Test (SAT) and California Achievement Test (CAT) questions, called the Point Biserial Correlation Coefficient (PBCC). This PBCC score measures the degree of correlation between correct answers to individual questions and total test scores, or in other words, whether people with high scores answer the question correctly and people with low scores answer the question incorrectly.

Using the Point Biserial Correlation Coefficient (PBCC), an item analysis of the 150 questions administered to 2,430 applicants identified by ethnic group from July 1982 through March 1983 indicates that 58 of those questions should not be used to screen applicants. Nearly two-thirds (64%) of the Reading section of the ATPE appears invalid while almost a third (32%) of the questions on the Math and one-fifth (20%) of the questions on the Grammar sections appear invalid. This means there are 58 items of questionable validity on a test where only 30 errors means the irrevocable loss of a career.

According to Dr. Hal Naumoff, Chairman of the Psychology Department at Phoenix College and a national consultant on employee selection and evaluation tests for private corporations, a test question should have a generally accepted minimum PBCC score of .33 or higher to be considered valid. However, as pointed out by Dr. David Krus of Arizona State University's Testing Services, a PBCC of .33 means that a person's knowledge accounts for only 10% of the reason a person got a question right or wrong. In other words, if the results of ATPE test sections are believed to measure knowledge, based on the answers given to the test questions, then a PBCC of .33 for a question means that knowledge only predicts 10% of

whether that person will answer that question correctly.

The 58 questions determined to be of questionable validity in this report all had PBCC scores below the .33 minimum calculated by comparing final math section scores for math questions, final grammar section scores for grammar questions, and final reading section scores for reading questions. This means questions were only considered invalid if knowledge of that section provided less than a 10% advantage to the person answering the question.

Technically, there are three types of test validity: content validity, construct validity, and discriminant validity. Content validity was the primary concern of the Department of Education in developing the ATPE, and consists primarily of determining whether the individual question items represent the topic that is to be tested. Construct validity refers to whether the questions measure what you intend the test to measure: in this case, a knowledge of basic skills needed in a classroom. Discriminant validity means that when a test divides a group into two subgroups (pass and fail), the members of each subgroup really belong in that group.

The Point Biserial Correlation Coefficient (PBCC) measures discriminant validity without concern for content. In other words, the PBCC measures whether the questions in the test measure what the test measures, regardless of what the test is supposed to measure. Every question written can be subject to different unintended interpretations that destroy its discriminant validity. The only clue to discriminant validity is consistency: does each question separate the group into the same subgroups that the overall test creates? Thus even though ATRA did not actually evaluate or even see the wording of the questions or answers, there are certain statistical re-

quirements that a test should meet.

For example, if a question is so difficult that 95% of the people answer it wrong, the 5% who answered it correctly should be among the most knowledgeable. Similarly, if a question is so easy that 95% of the people get it right, the 5% who missed the question should be the less knowledgeable. Actually, the level of difficulty makes no difference. If a test is supposed to measure something based on the answers given to the test questions, then the score a person receives on that test should predict whether that person can answer those questions correctly. If a test is supposed to measure knowledge of mathematics, then those who answer the math questions correctly should have much higher average math scores than those who answer it incorrectly.

For this reason, discriminant validity can be measured by comparing the final scores of those who answer a question correctly to the final scores of those who answered it incorrectly. Even though on any individual question, some people will guess correctly when they are not really knowledgeable, and some will accidentally answer incorrectly when they really are knowledgeable, the PBCC formula corrects for this and with a large number of people and a large number of questions these chance events will average out and disappear. For statistical purposes a large number is generally considered any number above 30. The ATPE has 50 questions in each of three sections, and over 2,400 tests were analyzed, making each section large enough for statistical evaluation of validity.

It is possible on questions where only a small number of people miss the question that inadvertent errors will account for most of those wrong answers and thus PBCC scores will be low. Even though the PBCC formula compensates for smaller numbers, it is still possible that valid-

ity scores will be inaccurate on extremely easy questions. For example, on the five Math questions that were answered correctly more than 97% of the time, all had PBCC scores below .29 indicating incorrect answers were probably due to inadvertent errors.

Five other Math questions answered correctly more than 95% but less than 97% of the time showed mixed results. Three had PBCC scores above .37, one was .329, and the third was substantially lower. No Grammar question was answered correctly more than 97% of the time, but of the four questions on the Grammar test that over 95% answered correctly, the PBCC scores were above .64 for three and above .53 for the fourth. On the Reading test, the same four questions were answered correctly over 95% and 97% of the time, with the highest PBCC score at .323 and the others lower.

Altogether, then, there were 18 questions answered correctly over 95% of the time on all three sections of the test, and only those nine that were answered correctly over 97% of the time were counted as invalid questions. Even if those nine questions are counted as valid, there are still 49 invalid questions remaining on the ATPE. These nine easy questions should be considered invalid, however, since they serve little use in a screening test if the only people who get them wrong do so because of inadvertent errors.

These PBCC statistics were determined for the entire group of 2,430 ethnically identified applicants who took the test during the first nine months of fiscal year 1982-83, including 1,840 identified ethnically as Whites, 161 identified as ethnic Hispanics, 52 identified as Black, 100 identified as Indian, 11 identified as Asian, and 266 identified as persons of mixed ethnic background who were not seeking teacher certification. The latter were primarily persons planning to teach at home. The scores did not include another

66 cases where ethnic identifiers were missing. The ATRA analysis validated each question by the score received on each subsection. The ATPE is actually scored on the overall test of all three subsections.

Based on these whole test scores, the ATPE screens out minority teacher candidates at significantly higher rates than Caucasians. According to a Department of Education report covering six of the nine months studied, 70% of Caucasians, 43% of Hispanics, and 23% of Blacks and Native Americans passed the test. Because of the small numbers involved, ATRA only examined two ethnic groups for this study: Caucasians and Hispanics. The ATRA analysis of the ATPE does not indicate that it is culturally biased. Only four additional questions would be considered seriously biased if only Hispanic scores were used for the Point Biserial calculations. Since subgroups decompose an average, subgroups are expected to have PBCC scores below any overall average. A measure of "seriously biased" would therefore be a PBCC below the .20 level. The CAT and SAT tests reject a question if an ethnic subgroup has a PBCC below the .20 level.

However, the ATPE illustrates the problem that invalid test items can have on test results. Peter Airasian, Thomas Kellaghan, and George Madaus commented in their critique of CONCEPTS OF SCHOOL EFFECTIVENESS (Nov, 1979) that:

> "If pupils cannot rely upon school specific instruction to provide a correct answer, they must rely on non-school experiences. As soon as pupils do this, measured achievement tends to become highly loaded on non-school, home background and social class factors." (page 24)

When a question is invalid the test taker must rely on something other than knowledge to

make a response. Subtle clues, cliches, word patterns, or cultural biases may result in the invalid question discriminating against certain groups. This tends to explain why PBCC scores on the ATPE when calculated using overall test scores instead of skill subsection scores tend to be higher when they should be lower. Normally knowledge of Grammar would not correlate well with Math questions, but if the questions are invalid they will be measuring some trait other than knowledge that the overall PBCC score would detect.

If language difficulties were responsible for failing the ethnic groups, one would expect spelling and word problems to be more objective in revealing this. Yet, only one of the ten questions in the spelling section of the Grammar test had a PBCC less than .33 in total or for Caucasians and Hispanics, and this question was missed primarily by Caucasians. The Hispanic PBCC on this spelling question was .36 even though over half the Hispanics missed this question. Similarly, of the five word problems on the math test, none had overall PBCC scores below .33, and the Hispanic PBCC scores were all above .25 with two above the .33 level.

At least for Hispanics, the ATPE does not appear to be invalid because of language problems. It is significant that the two most objective sections of the test, Math and Grammar, had the fewest invalid questions. Grammar imposes specific rules of punctuation and spelling that are easily determined. The Grammar section of the PBCC had 14 questions with PBCC scores above the .50 level. The Math section, including word problems and interpreting charts and graphs, did not have questions above the .50 level but it did have 20 questions above the .40 level. In contrast, the more subjective Reading section had only 4 questions with PBCC scores above the .40 level, and 32

of it's 50 questions were below the .33 level.

There are many reasons why seemingly simple test questions can be invalid. The validity of a question can depend very heavily on how well the question explains itself. Four of the 16 invalid questions on the Math section concerned interpreting graphs or charts. In fact, every question on maps, charts, or graphs in both the Math and Reading sections (9 questions in total) had PBCC scores less than .33 for the total group and for both Caucasians and Hispanics individually. Since these problems involve more subjective interpretations, this would tend to indicate that perhaps those questions are not adequately explained, have plausible distractors, or have other flaws.

The fact that questions which have very clear right and wrong answers, such as spelling and punctuation, also have very high PBCC scores on the ATPE emphasizes that validity can be achieved. The fact that questions with subjective interpretations have the lowest PBCC scores emphasizes the point that poor test construction is the likely culprit. Since seven of the 10 spelling questions had PBCC scores above the .50 level while all of the map and graph questions were below the .33 level, one is led to conclude that more specific and carefully constructed questions on maps and graphs would result in higher validity scores.

The failure to meet DISCRIMINANT validity standards is only one problem with the ATPE, albeit the most serious. The ATPE also has serious problems in CONSTRUCT validity. Although the ATPE is supposed to screen out teaching candidates deficient in the basic skills of math, grammar, and reading, the ATPE is graded as a whole. Thus even candidates who currently pass the ATPE may very well be deficient in one or more basic skills provided their knowledge in the other skills was sufficiently high to mask the deficiency. Therefore the construct ("concept" colloquially) of

screening out persons deficient in the basic skills is violated by the use of overall scores to determine pass or fail.

Even if the ATPE does achieve discriminant validity in the future, and even if the passing score were switched to an 80% score on each section, the test still lacks adequate construct validity. The cutoff score of 80% for passing the ATPE is a raw score. Because it is a raw score, it means that the questions that make up the test necessarily will be answered correctly by the great majority of those who take the test. This contributes to the difficulty of creating a valid test. People normally make some inadvertent errors. Consequently, a test should have a large enough number of difficult questions to ensure that inadvertent errors do not comprise a large percentage of the questions missed. The raw score on a more difficult test can always then be converted to some final score with a defined cutoff.

The difficulty level of the questions is really not the main problem with construct validity, it only contributes to the problem. Even if one accepts that most of the questions are answered correctly by the great majority of the examinees, one problem still remains: what represents an adequate level of knowledge to function in the classroom? The ATPE cutoff score of 80% does not appear to have any relation to any determined level of proficiency. The cutoff appears to have been selected primarily upon the political effect different cutoff scores would have on minorities. There is no evidence any attempt was made to find out what minimum level of proficiency was needed, let alone whether the test validly measured this proficiency.

The validity of the ATPE therefore suffers because many of the questions are not valid, the test is scored for overall proficiency rather than proficiency on the individual basic skills, the

score itself is a raw score that prevents using many difficult questions, and the cutoff point for passing or failing the exam was arbitrarily chosen. In terms of fulfilling the legislative mandate to screen teaching candidates for basic skill proficiency, this is equivalent to saying that firing a cannon suffices for hitting the target, when the cannon has neither the range nor accuracy to reach the target. The noise and smoke that attends the State Board's fulfillment of the ATPE requirement is full of sound and fury, but signifies nothing.

The validity of a screening test would seem to be the paramount consideration in its adoption. The ATPE was apparently adopted without concern for validity. This may well be due to the haste involved in developing the ATPE. In its literature and discussion of the ATPE during its adoption hearings the concept of validity was not addressed other than to note it purchased the "already validated, bias-free" ATPE from a private testing company. The primary concern during adoption was the difficulty level without regard to the validity and reliability of the test. The Department report titled SYNOPSIS OF ORIGIN AND DEVELOPMENT - ARIZONA TEACHER PROFICIENCY EXAMINATION notes in its opening paragraph:

> "Because of the timing of the legislation, the State Board had essentially a three-month timeline for implementation; in other states with similar requirements, a 2-3 year timeline has been normal procedure."

The State Board did field test the ATPE from November, 1980 to June, 1981, to develop a passing score level. However, when asked how the level of proficiency and validity was determined by the State Board, Associate Superintendent Dr. Tom Reno answered, "I'm not so sure there is any statistical reasoning to any of this. The decision re-

sulted from collective thinking and debate."

This collective thinking and debate by the State Board should have included some consideration of the capabilities and limitations of the accuracy of the test. One cannot expect to hit a target with a loose cannon, particularly if one doesn't aim the cannon. It appears the State Board was not adequately informed of the limitations of psychometric instruments, and did not fully comprehend the correct utilization of these selection tools.

In most testing situations a test is constrained by two important considerations. First, most tests are advisory only. Indeed, most test authorities are vociferous in warning against using tests in place of judgement. Second, most tests are norm referenced, meaning the scores represent a measure of how the examinee compares to other examinees. The ATPE is atypical in both of these regards.

First, the ATPE has more than an advisory function, it has a decision function. To become a teacher in Arizona requires passing the ATPE, the ATPE therefore has irrevocable decision power over certification. In this circumstance a much higher level of validity should be required of the test than of a test used simply in an advisory capacity.

Second, it is used as a criterion referenced test, meaning a score is supposed to indicate a level of knowledge rather than a comparison of relative knowledge. Norm referenced tests can tolerate lower validity scores because this invalidity applies to all examinees. It is much like playing a sport in the rain; rain falls on both teams equally, but their performance should not be considered a measure of ability. A norm referenced test with low validity has a "weak footing" that applies to all, it does not greatly alter the

relative positions of the examinees. A criterion referenced test cannot tolerate this flexibility.

It is not simply a matter of screening out candidates unfairly. The lack of adequate validity means both that persons who pass the test and persons who fail the test are labeled without valid reason. Simply put, a norm referenced test roughly indicates which of the candidates are more knowledgeable than other candidates. It can then serve as a guide when considering other factors. When the test is the only criterion involved, and when the questions are only tenuously connected to the knowledge they supposedly measure, the decision that is made by the test becomes arbitrary.

The achievement of high validity may be too expensive or difficult to expect of the ATPE. In that circumstance consideration should be given to providing more flexibility in interpreting the scores. For example, the test might be used in conjunction with the teacher residency program currently undergoing pilot testing in Arizona. Alternative methods of scoring the test might also be considered, such as norm referenced scores, percentile ranks, or standard deviations. These would have the effect of selecting the most proficient of the candidates without necessarily implying some level of proficiency.

The use of a test which only vaguely measures a knowledge of poorly defined basic skills at a level of proficiency that was never determined, may not have any meaning even if norm referenced. The field of educational measurement has developed a large body of knowledge, complete with caveats and protocols that should be adhered to when implementing this teacher evaluation program. Until the ATPE achieves a reasonable level of validity it should be discontinued.

(end)

****

TECHNICAL APPENDIX

The actual PBCC scores by ethnic group are provided in the following pages. In addition, charts showing these PBCC scores with their corresponding difficulty levels are also provided. These charts and graphs provide a visual display of the problems in the ATPE. From these materials it can be seen that more than 58 questions should be corrected. The .33 minimum is a cutoff, questions close to this minimum should be investigated. It should be remembered that the elimination of invalid questions will raise the PBCC scores of the remaining questions because final scores will become more accurate. Thus PBCC scores well above .33 should be the goal of test design. Only a few questions that are needed for content validity should be found near the minimum.

In examining the graphs, intuitively one would expect invalid questions to appear more difficult because answers would occur at random. This would show up on a graph of the test as a cluster sloping upward from the lower left to the upper right, meaning less valid questions with low PBCC scores would result in fewer correct answers. On the other hand, one would expect intuitively that valid questions would produce the reverse: the harder the question the more knowledge it would take to answer it. Thus the graph of a valid test would be a cluster sloping down from the upper left to the lower right, meaning the validity of test questions declines as they get easier because less knowledgeable people may know the answer and inadvertent errors by the more knowledgeable.

The graph of the Grammar section of the test shows the upward sloping tendency, indicating the

more difficult questions are actually just less valid, while the graph of the Math section appears the reverse. A reverse pattern like that indicates that the harder questions simply discriminate better than easier questions, requiring an incremental level of skill. The Reading section shows no pattern, however, indicating a random association between answers and validity.

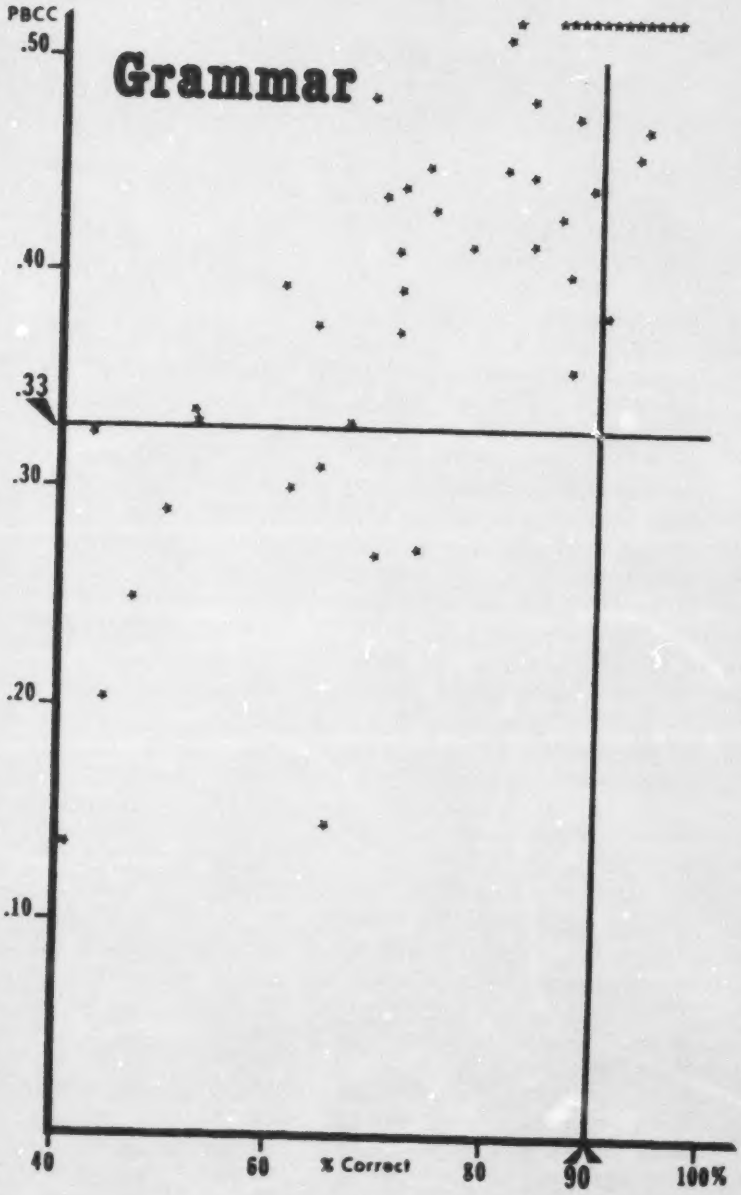
The graphs of PBCC scores against the percent correctly answered are divided by lines showing the .33 level of the PBCC and the 90% correct level. The graphs reveal that most invalid questions (below the horizontal line) were answered correctly over 90% of the time (to the right of the vertical line), as were many of the valid questions. Thirty of the 58 questions found invalid were answered correctly over 90% of the time (52%). This means that many very knowledgeable people were selecting some other answer as being correct. This could indicate that other plausible answers were present in the distractors, or that wording was unclear.

GRAMMAR SECTION
Point Biserial Correlation Coefficients
(multiplied by 100)

Question	Topic	All	White	Hispanic
1.	Question mark	47.05	39.87	36.41
2.	Exclamation mark	43.95	37.40	32.09
3.	Comma	41.22	39.16	31.53
4.	Comma	35.79	30.09	32.43
5.	Comma	40.01	34.84	23.11
6.	Italics	14.54	14.30	2.76
7.	Period	27.31	25.86	18.11
8.	Hyphen	31.20	29.34	25.57
9.	Underline	32.89	30.57	25.90
10.	Apostrophe	43.12	39.98	49.57
11.	Apostrophe	41.36	39.44	37.55
12.	Quotation marks	33.78	33.91	21.06
13.	Colon	68.67	62.49	73.53
14.	Colon	48.11	45.65	42.22
15.	Semicolon	48.19	46.31	39.79
16.	Capitals	41.08	42.25	26.08
17.	Capitals	47.69	43.00	41.79
18.	Capitals	53.67	48.36	56.74
19.	Double negative	42.38	37.24	33.73
20.	Subj.-verb agree	26.93	25.76	18.26
21.	Subj.-verb agree	44.87	39.33	49.91
22.	Verb Tense	44.87	41.88	43.98
23.	Pronoun	45.71	38.31	45.57
24.	Pronoun	39.21	38.42	39.91
25.	Pronoun	33.61	33.32	40.00
26.	Adverb	38.18	29.99	30.45
27.	Adjective	37.70	35.75	34.12
28.	Adjective	33.14	32.54	21.82
29.	Complete Sentence	65.44	57.93	70.66
30.	Complete Sentence	65.38	58.64	70.86
31.	Complete Sentence	56.74	49.91	57.77
32.	Complete Sentence	33.15	30.89	27.82
33.	Subordinate clause	29.06	27.12	28.31
34.	Inverted Sentence	13.83	15.92	15.53
35.	Detail & reason	30.71	26.17	28.19

36.	Detail & reason	44.45	37.31	55.69
37.	Order of detail	51.91	45.00	57.76
38.	Order of detail	24.85	20.63	31.12
39.	Main idea	44.00	39.03	44.40
40.	Main idea	64.41	56.59	72.01
41.	Spelling	51.24	47.79	66.67
42.	Spelling	58.08	52.36	70.08
43.	Spelling	53.55	47.04	63.61
44.	Spelling	39.50	40.04	44.03
45.	Spelling	59.04	52.89	52.62
46.	Spelling	61.59	55.16	69.45
47.	Spelling	53.00	49.87	56.61
48.	Spelling	37.38	35.66	41.57
49.	Spelling	52.04	45.13	42.62
50.	Spelling	20.32	20.66	36.08

The Grammar section of the ATPE has some validity, but ten, and possibly 14, of its questions should be rewritten or replaced with more valid questions. The upward slope indicates that wrong answers are associated with invalid questions rather than increasing difficulty. The band of .ars at the upper right symbolizes all questions above that level. Notice these easy Grammar questions are very good questions: those people who miss the question tend very strongly to lack knowledge in grammar, indicating the easier questions are very good discriminators.

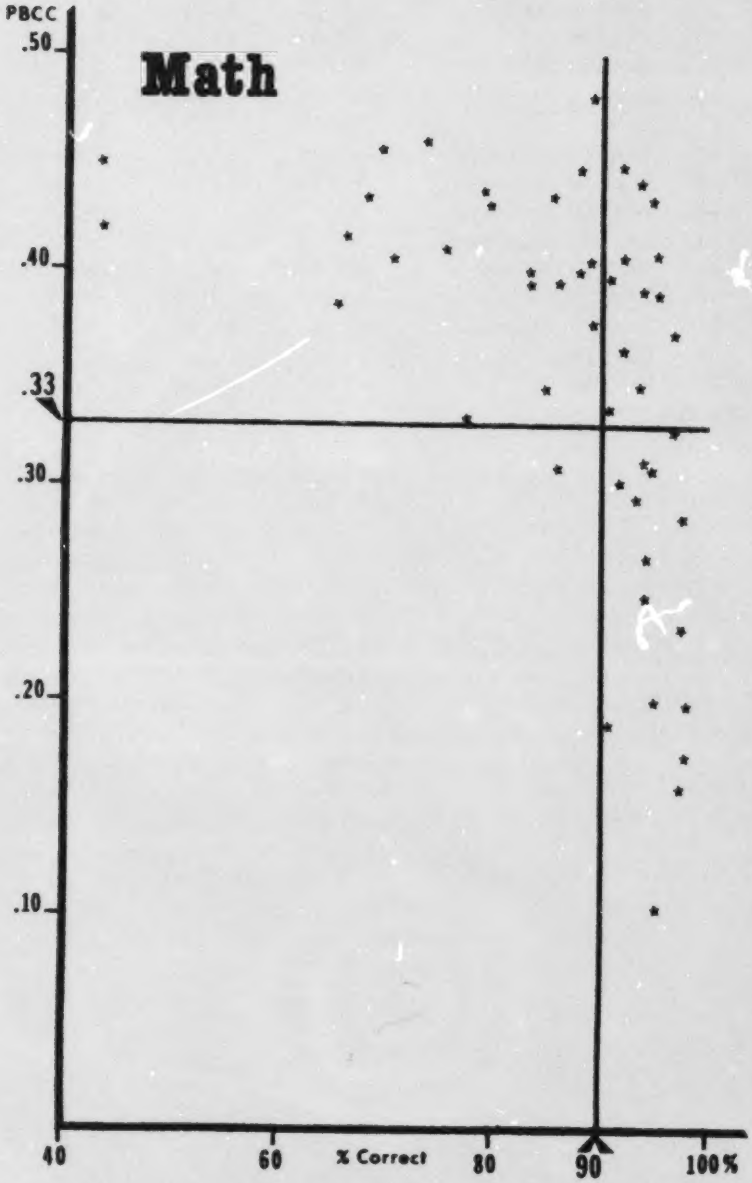

PBCC
.50
Grammar
.40
.33
.30
.20
.10
40
60
% Correct
80
90
100%

MATH SECTION
Point Biserial Correlation Coefficients
(multiplied by 100)

Question	Topic	All	White	Hispanic
1.	Addition-integers	17.66	12.45	2.98
2.	Addition-integers	16.07	15.51	18.25
3.	Addition-fractions	44.22	38.26	43.33
4.	Addition-fractions	44.63	40.95	48.93
5.	Addition-decimals	32.93	27.05	46.32
6.	Addition-decimals	19.01	18.07	23.73
7.	Subtract-integers	10.56	13.84	13.72
8.	Subtract-integers	20.09	18.07	27.21
9.	Subtract-fractions	40.27	35.63	40.44
10.	Subtract-fractions	43.58	39.27	40.70
11.	Subtract-decimals	37.10	35.28	49.30
12.	Subtract-decimals	40.59	41.06	37.71
13.	Multiply-integers	34.52	31.57	25.97
14.	Multiply-integers	31.30	25.83	19.84
15.	Multiply-decimals	34.82	27.15	36.43
16.	Multiply-decimals	40.10	34.96	34.70
17.	Division-integers	33.73	34.11	24.83
18.	Division-integers	29.47	25.30	12.61
19.	Division-decimals	26.67	20.43	21.67
20.	Division-decimals	30.88	26.27	25.89
21.	Fraction to %	28.48	19.98	13.60
22.	Fraction to %	40.51	41.07	36.85
23.	Decimal to %	43.82	43.75	40.55
24.	Decimal to %	30.73	31.60	39.82
25.	Interpret graph	23.70	15.01	0.16
26.	Interpret graph	30.20	25.03	23.25
27.	Interpret graph	24.97	18.02	13.60
28.	Interpret graph	19.97	11.82	8.77
29.	Word problem	38.38	38.88	26.77
30.	Word problem	39.70	35.80	25.04
31.	Word problem	44.74	38.82	36.38
32.	Word problem	46.00	44.56	27.91
33.	Word problem	43.11	38.74	43.60
34.	Averaging	39.04	30.07	38.52
35.	Averaging	39.16	28.47	27.48

36.	Averaging	36.27	29.65	36.83
37.	Averaging	43.21	41.74	38.84
38.	Perimeter	39.90	37.72	46.26
39.	Perimeter	47.72	42.35	36.95
40.	Area	37.57	31.05	19.24
41.	Area	41.97	41.70	33.10
42.	Liquid measure	45.00	45.32	41.69
43.	Liquid measure	40.55	31.19	33.31
44.	Liquid measure	39.55	37.27	24.29
45.	Liquid measure	39.80	34.33	29.97
46.	Time problem	33.20	29.10	14.31
47.	Time problem	45.72	45.94	23.40
48.	Time problem	43.50	39.84	40.27
49.	Metric	41.49	39.70	38.42
50.	Metric	40.93	37.92	36.35

The Math section seems to suffer from a combination of invalid questions and poor test design. There appears to be a sharp dichotomy between valid questions that form the downward sloping pattern of valid questions, and the vertical cluster of invalid questions. Most invalid questions are in the over 90% correct zone indicating many very knowledgeable people are missing those questions. This could indicate questions are not explained adequately, or difficulties arise in matching computations to answers. Clearly the more difficult questions are measuring a skill in math, but sixteen questions (and possibly 23), including 5 answered over 97% correctly, appear to have insufficient validity for a screening test.


PBCC
.50
Math
.40
.33
.30
.20
.10
40
60
% Correct
80
90
100%

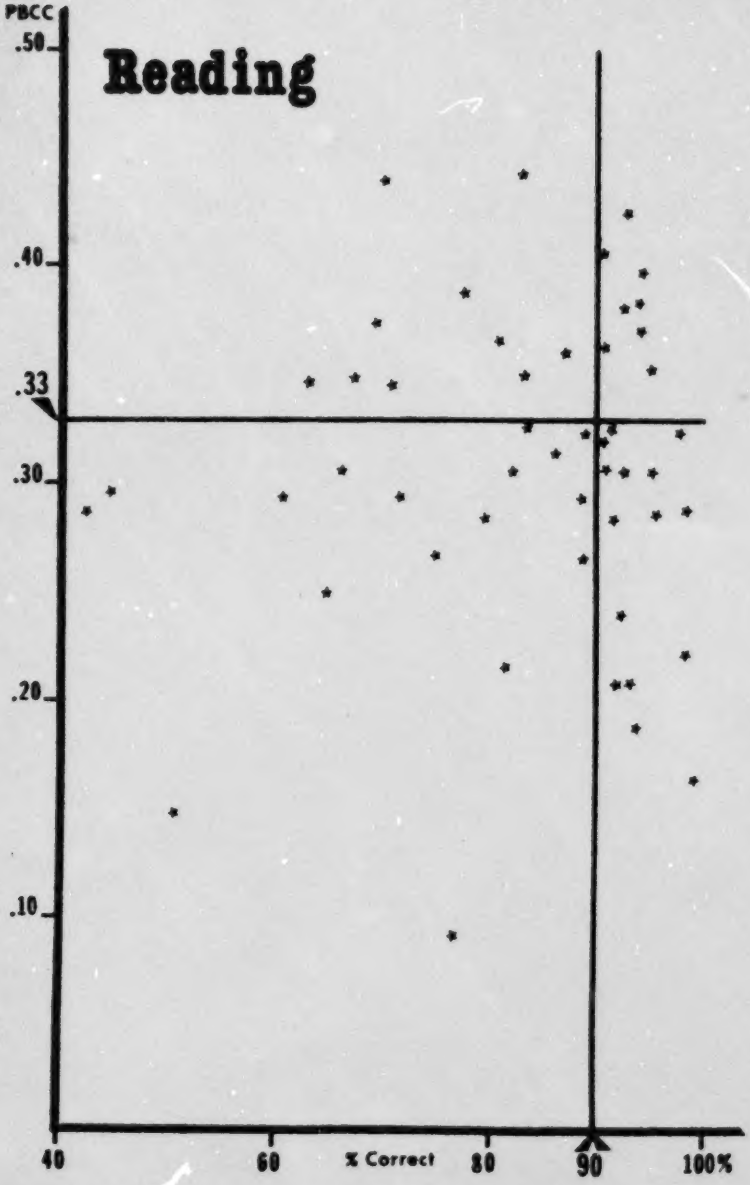
READING SECTION
Point Biserial Correlation Coefficients
(multiplied by 100)

Question	Topic	All	White	Hispanic
1.	Main idea	28.74	26.83	14.77
2.	Main idea	28.70	27.16	34.89
3.	Main idea	20.85	19.04	14.84
4.	Subordinate detail	35.52	34.63	33.19
5.	Subordinate detail	24.14	22.24	25.68
6.	Subordinate detail	37.95	36.29	45.62
7.	Cause & effect	36.37	32.37	33.37
8.	Cause & effect	35.88	34.03	31.12
9.	Sequence	43.89	42.68	27.43
10.	Sequence	38.88	36.15	40.95
11.	Sequence	32.65	34.97	16.29
12.	Inferred main idea	36.53	33.88	35.38
13.	Inferred main idea	35.17	28.75	32.19
14.	Inferred relations	37.18	34.49	28.34
15.	Inferred relations	40.65	37.66	33.27
16.	Inferred relations	39.84	37.50	31.21
17.	Predict outcome	44.16	41.88	31.41
18.	Predict outcome	24.75	23.06	10.60
19.	Predict outcome	29.70	29.37	26.66
20.	Figurative lang.	28.72	30.47	26.12
21.	Figurative lang.	37.54	37.92	31.24
22.	Figurative lang.	29.40	30.69	38.89
23.	Point of view	28.25	25.04	22.33
24.	Point of view	34.44	34.38	15.19
25.	Fact and opinion	34.87	34.17	32.04
26.	Fact and opinion	30.32	29.41	19.77
27.	Fact and opinion	31.62	26.73	34.19
28.	Generalizations	18.81	16.99	8.03
29.	Generalizations	30.83	30.41	11.40
30.	Generalizations	42.86	38.18	42.26
31.	Generalizations	34.79	31.64	39.64
32.	Contextual meaning	21.78	22.68	15.80
33.	Contextual meaning	32.35	30.67	27.66
34.	Contextual meaning	29.44	30.54	20.57
35.	Synonym	38.33	35.15	26.13

36.	Synonym	32.29	25.98	-n/a-
37.	Antonym	32.96	28.84	35.11
38.	Antonym	29.48	25.66	42.02
39.	Homophone	16.20	15.75	-2.03
40.	Alphabetize	22.18	21.82	11.85
41.	Alphabetize	20.85	20.11	3.57
42.	Dictionary skill	30.92	26.74	20.54
43.	Dictionary skill	28.28	24.88	7.13
44.	Table of contents	32.14	22.64	21.64
45.	Table of contents	30.83	26.67	25.90
46.	Map, Chart, Graph	26.82	25.03	24.31
47.	Map, Chart, Graph	9.27	9.80	14.34
48.	Map, Chart, Graph	30.77	26.83	17.62
49.	Map, Chart, Graph	14.96	16.65	22.41
50.	Map, Chart, Graph	26.48	26.57	18.57

The graph of the Reading section indicates severe problems with the validity of the test. Although a few questions seem to fall into the inadvertent error category, most show no pattern at all. It would appear from the graph that reading scores have no relation to the questions asked. In other words, the questions probably do not relate to reading ability. There is no indication at all of a reverse slope to indicate a measure of skill in reading. If anything, the graph would seem to indicate the pattern one would expect of a trivia test.

24


PBCC
.50
Reading
.40
.33
.30
.20
.10
40
60
% Correct
80
90
100%